Synthetic community Hi-C benchmarking provides a baseline for virus-host inferences

Rokaiya Nurani Shatadru[1,2], Natalie E. Solonenko[1,2], Christine L. Sun[2], Matthew B. Sullivan[1,2,3,4,#]

[1] Department of Microbiology, The Ohio State University.

[2] Center of Microbiome Science, The Ohio State University.

[3] Byrd Polar and Climate Research Center, The Ohio State University.

[4] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University.

[#] Corresponding author: sullivan.948@osu.edu; R914 Riffe Building, 496 W 12th Ave., Columbus, OH 43210

Running title: **Hi-C performance in virus-host linkage inferences**

Keywords: Hi-C, Virus-Host Interactions, Genomics

Abstract

Microbiomes are now recognized as key influencers of diverse ecosystems, but it is increasingly evident that viruses impose significant constraints on these microbial communities. While viromics has expanded virus genomic catalogs, identifying hosts for these viruses remains a major challenge due to the limitations in scaling for cultivation and to the uncertain reliability of *in silico* predictions for understudied virosphere regions. A promising recent advance, Hi-C, a proximity ligation-based method, aims to infer virus-host linkages by analyzing sequences from cross-linked virus and host genomic fragments. This approach has been applied in at least seven studies, yet its accuracy has not been systematically assessed. Here we evaluate Hi-C performance in predicting virus-host interactions using a synthetic community consisting of four bacterial strains and nine phages with known, experimentally determined, quantitative interactions. Our analysis revealed that Hi-C linkage scores used in the literature perform poorly (13% specificity, 100% sensitivity). By converting linkage scores to Z-scores and applying filtering (Z-score ≥ 0.5), we dramatically increased prediction accuracy, though at reduced sensitivity (96% specificity, 57% sensitivity). These findings provide empirical data and establish guidelines for interpreting Hi-C inferred virus-host linkages, with the aim of improving its reliability across diverse ecosystems.

Introduction

The microbiome of an ecosystem, which encompasses a diverse community of microorganisms, plays a crucial role in maintaining ecosystem stability (1). The complex network of interactions within the microbiome underpins essential ecological processes, including global phenomena like biogeochemical processes. Viruses are the key modulators of microbial ecosystems, regulating population dynamics and metabolic functions within the microbiome (2,3). Despite their importance, understanding virus-host interactions remains a major challenge in microbiome research.

Recent advances in viral metagenomics have led to the cataloging over 15 million viruses; however, the identification of their hosts have lagged significantly, with only 7% of virus-host relationships established to date (4). Further, among these identified virus-host interactions, only 11% have been experimentally validated, often reliant on the availability of pure cultures (4,5). While molecular techniques such as polymerase chain reaction (PCR) and probes can be effective, they require prior knowledge of viral sequences. Bioinformatic methods based on homology are generally reliable, but they often fail to detect novel viruses due to incomplete

reference databases (5). Although bioinformatic tools such as iPHoP (6), an aggregate of multiple bioinformatic methods, represent a promising approach for virus-host association prediction, they remain largely untested experimentally.

High-throughput Chromosome Conformation Capture (Hi-C), a proximity ligation method, is increasingly used for host identification in viral ecology across various ecosystems (7–14). In theory, Hi-C captures active viral infections and identifies other mobile genetic elements, such as plasmids, in the host cell. Briefly, Hi-C involves cross-linking bacterial and viral DNA in proximity using formaldehyde, followed by fragmentation with restriction enzymes (reviewed in (15)). The fragmented DNA is ligated, generating chimeric DNA composed of both viral and host fragments, which are resolved through high-throughput sequencing and bioinformatic analysis. Hi-C has been applied to identify hosts for many viruses by linking them to metagenome-assembled genomes (MAGs) in host associated ecosystems, such as fecal samples, and environmental ecosystems, like soil and hydrothermal mats (7,9,11,13).

Despite the growing use of Hi-C as a method for identifying virus-host interactions, several challenges remain in interpreting these linkages. First, while Hi-C has been experimentally validated for identifying plasmids in *Lactobacillus brevis* (16), no study to date has explored virus-host linkages based on confirmed relationships to establish proof of concept. Second, there is no benchmarking study assessing the sensitivity and specificity of Hi-C in understanding virus-host relationships. Finally, the proportion of spurious linkages remains unknown, and there is a lack of consensus on the best methods for filtering these linkages.

In this study, we employed a synthetic community with known virus-host interactions to assess the performance of Hi-C proximity ligation in reliably linking viruses with their hosts and evaluate its performance. Additionally, we also tested a sample-independent filtering method to reduce spurious connections and improve the accuracy of virus-host identification.

Materials and Methods

Strains and growth conditions

Four bacterial strains were used (**Table S1**): *Cellulophaga baltica* strain 18 was isolated from the Baltic Sea in 2000 (hereafter CBA 18; (17)), while *Pseudoalteromonas* sp. H71, H105 and 13-15 were isolated from the North Sea in 1990 (hereafter PSA H71, PSA H105, and PSA 13-15; (18)). Nine bacteriophages were used: five infecting *C. baltica* strains and four infecting *Pseudoalteromonas* strains (**Table S1**). The CBA phages φST, isolated on CBA 3 (strain not

included in this study); φSM, isolated on CBA 3; φ46:1, isolated on CBA 46 (strain not included in this study); φ14:2, isolated on CBA 14 (strain not included in this study); and φ18:1, isolated on CBA 18, were isolated from the Baltic Sea between 2000 and 2005 (17). The PSA phages, PSA-HM1, isolated on H71; PSA-HS4 isolated on PSA H105; PSA-HP1, isolated on PSA H100 (strain not included in this study); and PSA-HS2, isolated on PSA 13-15, were isolated from the North Sea in 1990 (19).

Bacteria were grown on Pseudoalteromonas-Cellulophaga (PC) plates (20.5g Sigma Sea Salts, 1.5g peptone, 1.5g proteose peptone, 0.6g yeast extract, 13g agar/L) at room temperature (RT). Single colonies were inoculated and grown stationary at RT overnight in PC liquid growth medium (20.5g Sigma Sea Salts, 0.75g peptone, 0.5g yeast extract, 0.25g proteose peptone, 0.25g casamino acids, 1.5mL glycerol/L). Phage strains were stored in phage buffer (20.5g Sigma Sea Salts/L) and plaque forming units (PFUs) enumerated using the agar overlay method (20) with 3.5mL molten soft agar (20.5g Sigma Sea Salts and 6g low melting point agarose/L) and 300µl overnight bacterial culture per plate.

Adsorption kinetics experiments

Each phage–host pair was tested to determine which phages were able to adsorb to which host strains. Each bacterial strain was inoculated and transferred as for the synthetic community sampling experiments. Once each strain had reached mid-exponential phase and at least 1×10^8 CFU/mL, adsorption experiments were performed for each phage–host pair in duplicate. Cells and phage were incubated together in 1 mL PC medium at a multiplicity of infection of 1 (1×10^8 CFU/mL and 1×10^8 PFU/mL). A no bacteria control was also performed with each phage to ensure phage titer did not decrease due to other factors (e.g., sticking to the tube). For phages with no known host in the synthetic community (phages 14:2 and SM) or known poor adsorption to mock community bacteria (phage φ ST), their hosts of isolation were included as a positive control. Total and free phage were sampled and plated to determine PFU/mL at 0 min and 60 min, or 30 min where the known latent period is less than 60 min (phages HM1 and 46:1). Adsorption was determined by comparing the free phage concentration at 60 min or 30 min to that at 0 min. A t-test was performed and any result that showed a significant decrease in free phage over time was considered positive for adsorption for that phage–host pair.

Follow-up adsorption kinetics experiments (**Table S2**) were conducted on phage–host pairs showing significant adsorption in the initial experiment above. These experiments were performed in biological triplicates to reduce the chances of multiple adsorptions to a single host

110 cell, and sampled every few minutes for ~30 min, but otherwise similar to the above. Some experiments did not include a 15 min timepoint – for these, data from 16 min is given instead.

Synthetic community sample collection

One synthetic community was made, with three replicates. All phage concentrations were based on infectious particles only, either when infecting bacterial hosts contained in the synthetic
115 community or, for phages with no host in the synthetic community, on the host of isolation. Synthetic community 1 (SymCom 1) had equal numbers of each phage; final concentrations of each bacterial strain and phage are given in **Table S1**. Overnight cultures of each bacterial strain were transferred into fresh PC media and grown stationary at RT until mid-exponential phase and a concentration of $>1\times10^8$ colony forming units (CFU)/mL was reached. Concentration was
120 determined by reading the optical density of each culture at 600 nm and using a regression equation of OD vs CFU/mL. A total of 5×10^8 CFUs from each mid-exponential bacterial strain were combined in a 50 mL tube. The nine phages were then added to each tube at the multiplicity of infection (MOI) indicated in **Table S1**. After 15 minutes to allow phages to adsorb to their bacterial hosts, the community was diluted to 50 mL with PC phage buffer to slow adsorption and
125 give a final cell concentration of 1×10^7 CFU/mL per strain. Aliquots of 1 mL were sampled, and flash frozen in liquid nitrogen. Samples were stored at -80°C until processing.

Hi-C sample library preparation

Before Hi-C library preparation, samples were thawed, centrifuged to remove extracellular DNA, and the pellet was resuspended in the crosslinking solution provided in the ProxiMeta kit
130 (Phase Genomics). Hi-C libraries were prepped from the resuspended solution for each replicate of the synthetic community, following the manufacturer's instructions. Samples were pooled and sequencing was performed by SeqCenter (Pittsburg, PA, USA) on an Illumina NovaSeq X 10B 300cyc lane.

Hi-C data processing

135 The sequences from each sample were quality filtered using BBduk from the BBmap suite (21) to remove the adapters and PhiX reads from the raw reads, and to trim and filter out low quality reads. BWA mem (22) was used with the -5SP flag to align the Hi-C reads from each sample to the reference genomes of the members of the synthetic community. To normalize the Hi-C linkages, MetaCC (23) was used with the enzyme flags Sau3AI and MluCI. Further analysis
140 was done by utilizing the sparse matrix output 'Normalized_contact_matrix' from MetaCC by adding the contact scores for genomes split across multiple contigs, resulting in one contact score for each virus-host pair. Z-scores were calculated for each synthetic community replicate

according to the following equation where x is the contact score of a specific pair of host and virus, μ is the mean of the contact scores of all the host and virus linkages, and σ is the standard deviation of the contact scores of that sample.

$$Z = \frac{x - \mu}{\sigma}$$

The rates of sensitivity and specificity of the predictions were calculated according to the following equations:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \times 100$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \times 100$$

The scripts used for data preparation, analysis, and figure generation are available upon request.

Results and Discussion

Design of a synthetic community to evaluate Hi-C performance in virus-host linkage inference

To evaluate the accuracy of Hi-C in identifying virus-host interactions, we designed an experiment using a synthetic community (SynCom) composed of four well-characterized bacterial strains and nine phages with known relationships and efficiency of plating (12–14) (**Figure 1A, Figure S1, Table S1**). In this synthetic community, *Cellulophaga baltica* strain 18 (CBA 18) has three infecting phages, and three *Pseudoalteromonas* strains (PSA 13-15, PSA H105, PSA H71) were infected by 2, 1, and 1 phage strains, respectively. Moreover, there were two additional *Cellulophaga baltica p*hages that do not infect the strain present in the synthetic community. SynCom composition included bacterial host to phage ratios that varied (1:1, 1:2, and 1:3) and 'distractor' phages known not to infect at least one host to examine how varying infection dynamics influenced Hi-C performance. These quantitatively defined interactions formed the basis for all our subsequent Hi-C performance assessments.

To establish Hi-C as a method for identifying virus-host interactions, we first constructed "SynCom 1", where bacteria and infective phage particles were added at high concentrations (10^7 CFU/mL for bacteria and 10^6 PFU/mL for phages) (**Figure 1A, Table S1**). The multiplicity of infection (MOI) of 0.1 exceeds natural levels but was chosen to create a best-case Hi-C scenario while minimizing the occurrence of multiply-infected bacteria. Phage amounts were adjusted based on their infection efficiency (determined by plating), ensuring an equal quantity of infectious phages across all strains **(Figure S1).** After mixing the phages with bacteria, we incubated the

mixture for 15 minutes to allow adequate phage adsorption, as pairwise experiments showed 74-
175 99% adsorption within this time frame (**Table S2**). At the end of incubation, we prepared three technical replicates by aliquoting 1mL of the synthetic community, which were immediately flash-frozen to halt phage activity (**Figure 1B**). Hi-C libraries were then prepared using the PhaseGenomics ProxiMeta kit, and three SynCom 1 samples were sequenced on Illumina technology, yielding with an average depth of 385M total reads per sample (**Table S3**).

180 After pair-end Illumina sequencing, Hi-C samples were processed to bioinformatically investigate virus-host linkages (see Materials and Methods for details). Briefly, sequencing reads were aligned to bacterial and phage reference genomes. A 'linkage' is defined as a read pair that aligns (or 'connects') to sequences either within the same genome, or between different genomes. The majority of linkages (62% on average, **Table S4**) were from the same genome, consistent
185 with intracellular genome fragments remaining in close proximity, as previously reported (16). The linkages between different genomes (virus-virus, virus-host, or host-host) comprised 2% to 18% of all linkages. Although many Hi-C studies on virus-host interactions did not normalize linkages (**Table S5**), we consider normalization essential, as Hi-C linkages depend greatly on the number of restriction sites, length, and coverage of the reference genomes (15). To obtain normalized
190 linkages, we applied MetaCC (23), which uses a binomial regression approach to adjust the number of linkages between genomes and produce a normalized 'contact score'. This normalization, based on restriction site count, genome length, and read coverage, has been used in a virus-host interaction study of hydrothermal mats (11). All non-zero contact scores from MetaCC were interpreted as inferred linkages in this study.

195 *Detection of true and spurious phage-host linkages in the synthetic community*

We identified both expected and unexpected phage-host linkages in SynCom 1. For the expected linkages, we observed all seven known interactions identified in prior single phage-host infection assays (**Figure 2, Figure S1**), confirming high sensitivity with 100% true positive detection across three replicates (**Table S6**). However, we also observed unexpected linkages
200 involving phage-bacteria pairs thought to represent negative interactions in the prior single phage-bacterial infection assays (**Figure 2, Figure S1**). For example, φ14:2, which lacked a corresponding host in SynCom 1, showed linkages with multiple bacterial strains across all three replicates (**Figure 2B**). Specificity (true negative frequency) was quite low, averaging just 13% across the three replicates (**Table S6**). In terms of reproducibility, the presence/absence of
205 linkages was highly consistent between replicates; all 36 possible interactions were identical in replicates 1 and 2, whereas replicate 3 showed eight additional linkages that were absent in the others (**Figure 2B**).

We next evaluated the magnitude of the contact scores to assess its value as a proxy for virus-host linkage prediction confidence. Ideally, we would expect similar contact scores for the same linkages across replicates within SynCom 1. However, we observed unexpected and substantial differences in contact scores for both correct and incorrect linkages. For example, CBA 18 and φ18-1 (a known interaction) had contact scores of 171, 97, and 13 (normalized per 100,000 reads to account for sequencing depth differences) in replicates 1, 2, and 3, respectively. Similarly, CBA 18 and H71 (which do not have a known interaction), had contact scores of 41, 50, 6 (normalized per 100,000 reads), in replicates 1, 2, and 3, respectively. Despite these variations in contact scores for the same linkages across replicate, we observed that correct linkages generally had higher contact scores than incorrect ones (**Figure 2A,** black vs grey dots). In fact, 85% of the time, across all the replicates, the highest contact score for each phage was to its correct host **(Figure S2).** Nevertheless, the observed variability in contact score values present a challenge for establishing a universal threshold to minimize false positives.

To statistically evaluate the performance of contact scores as a measure of identifying virus-host interactions, we used receiver operating characteristics (ROC) curves to analyze all the interactions across replicates (**Figure S3**). The ROC curve provides a comprehensive view of the model's ability to distinguish between correct and incorrect linkages at various threshold settings. This analysis yielded an area under the curve (AUC) of 0.81, indicating that the contact scores effectively differentiate linkages (with AUC values closer to 1 representing better performance) (**Figure S3**). However, a key question remains: how can we reduce false positives, considering the scaling of the contact scores? To address this, we explored methods used in other studies to mitigate noise. We found that various strategies have been implemented to reduce noise, such as using inter-genome and intra-genome ratios or requiring a minimum number of linkages (e.g., 2 or 10) (9,11,13). However, these approaches were defined without benchmarking against known relationships, leaving uncertainty about the true impact of these cut-offs.

Evaluating the use of Z-scores to improve Hi-C virus-host linkage accuracy

We experimented with the use of Z-scores to standardize contact scores across samples, aiming to establish a potential filtering threshold. A Z-score is a widely used statistical metric that indicates how many standard deviations a specific value is from the mean. As a dimensionless, ratio-based metric, Z-scores provide a universal way to assess the relative magnitude of a value within a sample and are commonly used in RNA-seq analysis to identify differentially expressed genes (24). When we converted the contact scores to Z-scores for each sample, we found that most correct linkages had Z-scores above 0 (**Figure 2A**), which was encouraging. We then plotted the Z-scores from all SynCom 1 replicates in a ROC curve, yielding an AUC of 0.89 (**Figure S3**).

The increase of the AUC from 0.81 to 0.89 demonstrated that Z-scores served as a better discriminatory factor compared to the MetaCC contact scores for identifying correct virus-host linkages.

Applying a Z-score filter to the SynCom 1 Hi-C data resulted in a notable improvement in signal. Based on the ROC curve (**Figure S3**), we selected a conversative threshold of 0.5 to reduce noise in the dataset. This threshold was chosen because, according to the ROC curve, 0.5 yielded the highest specificity and highest sensitivity across all replicates. Applying this threshold removed 82% of virus-host linkage inferences, which reduced average sensitivity from 100% to 57%, but drastically increased the average specificity from 13% to 93% (**Table S6**). Since the goal was to minimize false positives and improve overall accuracy, this represents a significant improvement in predictive performance. This threshold showed consistent performance across all replicates in this study. These findings suggest that a standardized Z-score threshold could enhance Hi-C filtering methods in future analyses. However, further evaluation in biological replicates and additional other sample types is needed to determine whether this specific 0.5 cut-off is universally applicable or should be empirically defined for each sample set.

Conclusion

In this study, we presented the first proof-of-concept for using Hi-C proximity ligation to identify virus-host linkages by leveraging a synthetic community with known interactions. Our findings show that unfiltered Hi-C linkages lead to poor accuracy. However, when a Z-score threshold ($Z \geq 0.5$) was applied, performance improved drastically, enhancing both sensitivity and specificity. These insights are critical when designing Hi-C experiments with natural samples, in order to optimize and increase success of Hi-C-based host-virus identification. Overall, this study establishes Hi-C as a promising method for identifying virus-host linkages while highlighting opportunities for refinement.

Competing Interests

None

Acknowledgments

We thank Ben Bolduc and Garrett Smith for bioinformatic assistance. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-SC0023307 to M.B.S. This research is a

275 contribution of the EMERGE Biology Integration Institute, funded by the National Science Foundation, Biology Integration Institutes Program, Award #2022070. This work was also supported by NIH grant R01AI169865 subaward to M.B.S.

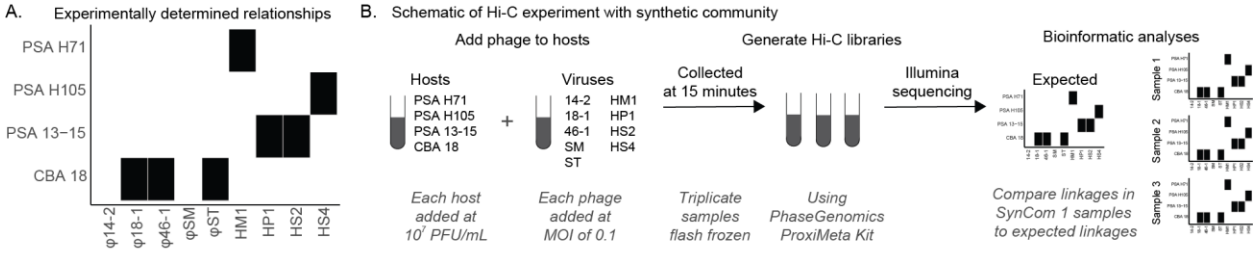

Figure 1. Synthetic communities and experimental schema used to assess Hi-C virus-host linkages. **A.** Synthetic communities were built from four bacterial strains (CBA = *Cellulophaga baltica*; PSA = *Pseudoalteromonas*) and 9 phages that were experimentally evaluated for infection in pairwise combinations via traditional plaque assays. Black boxes denote that the virus successfully plaques on the bacterial strain, whereas white (missing) boxes denote a negative, non-plaquing interaction. **B.** Schematic representation of the Hi-C experiment used to test virus-host relationships. After generating the synthetic community with the organisms mentioned above, Hi-C libraries were prepared and sequenced. Subsequently, bioinformatic analyses were performed to determine whether the expected virus-host linkages (denoted with black boxes) were observed.

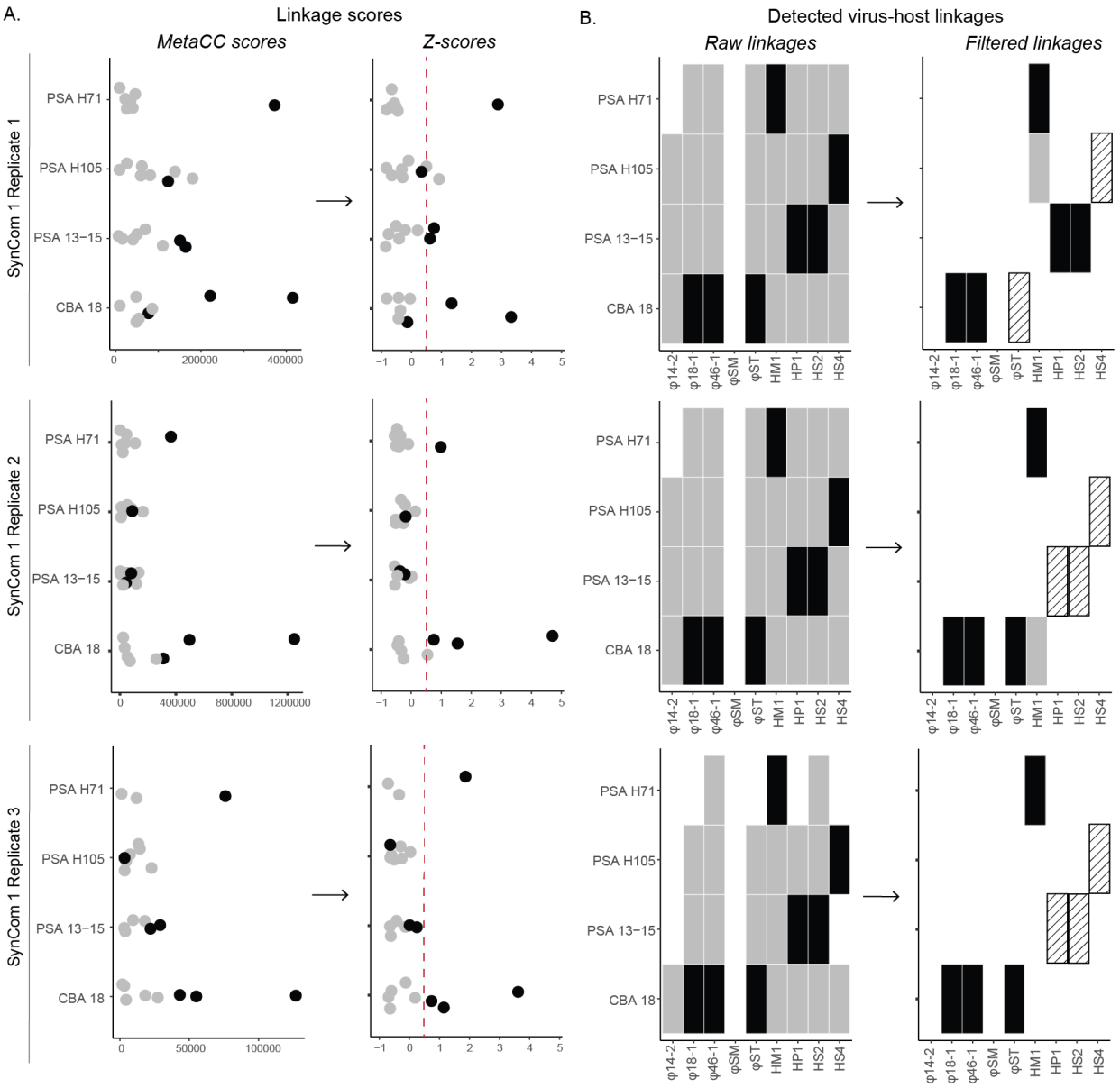

Figure 2. Hi-C linkages from SynCom 1. A. Contact scores (left) and Z-scores (right) calculated for each replicate of SynCom 1, categorized by host strains. The black dots indicate correct virus-host linkages, the grey dots indicate incorrect virus-host linkages. The red vertical dotted line is drawn at Z-score = 0.5. **B.** Virus-host linkages determined from the raw scoring method used in the literature (i.e., a non-zero contact MetaCC score; left) or using a Z-score approach (i.e., requiring a Z-score generated from MetaCC scores to be above 0.5; right). The black boxes denote correct linkages, gray boxes denote incorrect linkages, and the striped boxes denote correct linkages that were not detected.

References

1. Desai SA, Patel VP, Bhosle K, Nagare S, Thombare K, Hashem A, et al. Microbiome and Ecosystem Approaches. In: Parray JA, editor. Progress in Soil Microbiome Research [Internet]. Cham: Springer Nature Switzerland; 2024 [cited 2024 Dec 14]. p. 37–52. Available from: https://doi.org/10.1007/978-3-031-71487-0_2

2. Suttle CA. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol. 2007 Oct;5(10):801–12.

3. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018 Aug;3(8):870–80.

4. Camargo AP, Nayfach S, Chen IMA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. 2023 Jan 6;51(D1):D733–43.

5. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. Smith M, editor. FEMS Microbiol Rev. 2016 Mar;40(2):258–72.

6. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLOS Biol. 2023 Apr 21;21(4):e3002083.

7. Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, Press MO, et al. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. Genome Biol. 2019 Aug 2;20(1):153.

8. Marbouty M, Thierry A, Millot GA, Koszul R. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. Duerkop BA, Garrett WS, Duerkop BA, Quince C, editors. eLife. 2021 Feb 26;10:e60608.

9. Ivanova V, Chernevskaya E, Vasiluev P, Ivanov A, Tolstoganov I, Shafranskaya D, et al. Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients. Front Microbiol. 2021;12:770323.

10. Cuscó A, Pérez D, Viñes J, Fàbregas N, Francino O. Novel canine high-quality metagenome-assembled genomes, prophages and host-associated plasmids provided by long-read metagenomics together with Hi-C proximity ligation. Microb Genomics. 2022 Mar 17;8(3):000802.

11. Hwang Y, Roux S, Coclet C, Krause SJE, Girguis PR. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat Microbiol. 2023 Apr 6;1–12.

12. Cheng Z, Li X, Palomo A, Yang Q, Han L, Wu Z, et al. Virus impacted community adaptation in oligotrophic groundwater environment revealed by Hi-C coupled metagenomic and viromic study. J Hazard Mater. 2023 Sep 15;458:131944.

13. Wu R, Davison MR, Nelson WC, Smith ML, Lipton MS, Jansson JK, et al. Hi-C metagenome sequencing reveals soil phage–host interactions. Nat Commun. 2023 Nov 23;14(1):7666.

14. Marbouty M, Baudry L, Cournac A, Koszul R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. Sci Adv. 2017 Feb 17;3(2):e1602105.

15. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. Methods San Diego Calif. 2015 Jan;72:65–75.

16. Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ. 2014 May 27;2:e415.

17. Holmfeldt K, Middelboe M, Nybroe O, Riemann L. Large Variabilities in Host Strain Susceptibility and Phage Host Range Govern Interactions between Lytic Marine Phages and Their Flavobacterium Hosts. Appl Environ Microbiol. 2007 Nov;73(21):6730–9.

18. Wichels A, Biel SS, Gelderblom HR, Brinkhoff T, Muyzer G, Schütt C. Bacteriophage diversity in the North Sea. Appl Environ Microbiol. 1998 Nov;64(11):4128–33.

19. Duhaime MB, Wichels A, Sullivan MB. Six Pseudoalteromonas Strains Isolated from Surface Waters of Kabeltonne, Offshore Helgoland, North Sea. Genome Announc. 2016 Feb 11;4(1):e01697-15.

20. Sambrook J. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory; 2003. book.

21. BBMerge – Accurate paired shotgun read merging via overlap | PLOS ONE [Internet]. [cited 2023 Jul 12]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185056

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009 Jul 15;25(14):1754–60.

23. Du Y, Sun F. MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. Nat Commun. 2023 Oct 6;14(1):6231.

24. Kobayashi Y, Kushihara Y, Saito N, Yamaguchi S, Kakimi K. A novel scoring method based on RNA-Seq immunograms describing individual cancer-immunity interactions. Cancer Sci. 2020;111(11):4031–40.